My AI Wants to Know if This Will Be on the Exam: Testing OpenAI's Codex on CS2 Programming Exercises

Abstract

The introduction of OpenAI Codex sparked a surge of interest in the impact of generative AI models on computing education practices. Codex is also the underlying model for GitHub Copilot, a plugin which makes AI-generated code accessible to students through auto-completion in popular code editors. Research in this area, particularly on the educational implications, is nascent and has focused almost exclusively on introductory programming (or CS1) questions. Very recent work has shown that Codex performs considerably better on typical CS1 exam questions than most students. It is not clear, however, what Codex's limits are with regard to more complex programming assignments and exams. In this paper, we present results detailing how Codex performs on more advanced CS2 (data structures and algorithms) exam questions taken from past exams. We compare these results to those of students who took the same exams under normal conditions, demonstrating that Codex outscores most students. We consider the implications of such tools for the future of undergraduate computing education.